

EFFICIENT COMPUTATION OF THE JOINT SAMPLE FREQUENCY SPECTRA FOR MULTIPLE POPULATIONS

BY JOHN A. KAMM^{*}, JONATHAN TERHORST[†] AND
YUN S. SONG^{*,‡}

University of California, Berkeley

A wide range of studies in population genetics have employed the sample frequency spectrum (SFS), a summary statistic which describes the distribution of mutant alleles at a polymorphic site in a sample of DNA sequences. In particular, recently there has been growing interest in analyzing the joint SFS data from multiple populations to infer parameters of complex demographic histories, including variable population sizes, population split times, migration rates, admixture proportions, and so on. Although much methodological progress has been made, existing SFS-based inference methods suffer from numerical instability and high computational complexity when multiple populations are involved and the sample size is large. In this paper, we present new analytic formulas and algorithms that enable efficient computation of the expected joint SFS for multiple populations related by a complex demographic model with arbitrary population size histories (including piecewise exponential growth). Our results are implemented in a new software package called *momi* (MOran Models for Inference). Through an empirical study involving tens of populations, we demonstrate our improvements to numerical stability and computational complexity.

1. Introduction. The sample frequency spectrum (SFS) is the distribution of allele frequencies at a polymorphic site in a collection of DNA sequences randomly drawn from a population. This summary statistic is used in a variety of inference problems in population genetics [5, 11, 13, 16, 17, 18, 19, 22, 33, 34, 42], often in the context of likelihood-based analysis of single nucleotide polymorphism (SNP) data. Over the past several years, there has been much interest in analyzing the joint SFS data from multiple populations to infer complex demographic models involving population size changes, population splits, migration, and admixture. Inferring population

^{*}Supported in part by an NIH grant R01-GM109454.

[†]Supported in part by a Citadel Fellowship.

[‡]Supported in part by a Packard Fellowship for Science and Engineering, and a Miller Research Professorship.

AMS 2000 subject classifications: Primary 92D15; secondary 65C50, 92D10

Keywords and phrases: the coalescent, demographic inference, sum-product algorithm, fast Fourier transform

demographic histories is not only intrinsically interesting, for example in dating events such as the out-of-Africa migration of modern humans [19, 38], but is also important for biological applications, such as distinguishing between the effects of natural selection and demography [3, 6].

Likelihood-based inference methods using the SFS require accurate computation of the expected SFS under a given demographic model. As further detailed below, however, existing methods suffer from numerical instability and high computational complexity when multiple populations are involved and the sample size is large. The joint SFS for multiple populations describes the distribution of joint allele frequencies across the different populations. In this paper, we tackle the problem of computing the expected joint SFS for many populations, given a complex demographic model relating them.

The SFS has been studied in the context of two dual processes, the Wright-Fisher diffusion [25] and Kingman’s coalescent [15], and both approaches can be used to compute the multi-population SFS. In the diffusion approach of Gutenkunst et al. [19], which was later further extended [17, 31], one numerically solves partial differential equations forward in time to approximate the distribution of joint allele frequencies at present. The diffusion framework has the advantage of being applicable to arbitrary demographic models, but its computational complexity grows exponentially with the number of populations, and current implementations have difficulty handling more than three [19] or four populations [31].

In the coalescent approach, the SFS is computed by integrating over all genealogies underlying the sample. This can be done via Monte Carlo or analytically. Monte Carlo integration approach [34] can effectively handle arbitrary demographic histories with a large number of populations, and Excoffier et al. [13] have recently developed a useful implementation. However, when the number \mathcal{D} of populations (or demes) is moderate to large, most of the $O(n^{\mathcal{D}})$ SFS entries, where n denotes the sample size, will be unobserved in simulations, and thus the Monte Carlo integral may naively assign a probability of 0 to observed SNPs. Monte Carlo computation of the SFS thus requires careful regularization techniques to avoid degeneracy issues.

An alternative to the Monte Carlo approach is to compute the SFS exactly via analytic integration over coalescent genealogies [18, 42]. For a demography involving multiple populations, this can be done efficiently by a dynamic program [8, 9]. This algorithm is more complicated and less general than both the Monte Carlo and diffusion approaches: while it can handle population splits, merges, size changes, and instantaneous gene flow, it is difficult to include continuous gene flow between populations. However, it

scales well to a large number \mathcal{D} of populations, since it only computes entries of the SFS that are observed in the data, and ignores the $O(n^{\mathcal{D}})$ SFS entries that are not observed. Unfortunately, existing coalescent-based algorithms [8, 9, 42] do not scale well to a large sample size n , both in terms of running time and numerical stability. In particular, the algorithm relies on large alternating sums that explode with n and exhibit catastrophic cancellation.

In this paper, we significantly improve the computational complexity and numerical stability of the coalescent approach. We show how the alternating sums can be avoided altogether, and replaced with faster and more stable formulas. Moreover, we introduce a second speedup by replacing the coalescent with a Moran model in the dynamic program.

The dynamic program algorithm to compute the SFS involves splitting the demography into its component subpopulations, each of which contains a single population coalescent, but truncated at some time τ in the past. In Section 2, we focus on this *truncated coalescent*. In particular, we focus on computing the *truncated SFS* $f_n^\tau(k)$, the expected number of mutations arising in the time interval $[0, \tau)$ which subtend exactly k out of n individuals sampled at time 0. We give an algorithm for computing $f_n^\tau(k)$ efficiently, using recurrence relations combined with results from Polanski, Bobrowski and Kimmel [36], Polanski and Kimmel [37] and Bhaskar, Wang and Song [5]. We also provide an alternative formula for $f_n^\tau(k)$ based on *the coalescent with killing*.

In Section 3, we describe the coalescent algorithm of Chen [8, 9], and show how our formulas for $f_n^\tau(k)$ improve its computational complexity. For the special case where the demographic history forms a tree, we introduce an additional speedup by replacing the coalescent with a Moran model. For such tree-shaped demographies, we can compute the observed SFS entries in $O(n^2\mathcal{D} + n\log(n)DL)$, where n is the sample size, \mathcal{D} is the number of populations at the present, and L is the number of observed entries in the SFS. This is an improvement over the $O(n^5\mathcal{D} + n^4DL)$ complexity in Chen [8, 9]. For more general demographic histories with migration or admixture, the algorithm of Chen [8, 9] is $O(n^5V + WL)$, where V is the number populations (vertices) throughout the history, and W is a term that depends on n and the graph structure of the demography; we improve this to $O(n^2V + WL)$. In future work, we will give explicit expressions for W , and extend our Moran-based speedup to demographies with pulse migration.

We note that some of our improvements are related to results in Bryant et al. [7], whose $O(n^2\log(n)DL)$ algorithm computes the one-locus likelihood for species trees with recurrent mutation and piecewise constant population sizes. By contrast, our method, like that of Chen [8, 9], considers an infinite

sites model [26] without recurrent mutation, but can handle exponentially growing population sizes. In fact, our method goes even further, and easily accommodates arbitrary population size changes.

In Section 4, we demonstrate the improved speed and accuracy of our algorithm in a numerical study involving tens of populations. We implement and release our algorithm in a publicly available Python package, *moni* (MOran Models for Inference). Proofs of the mathematical results presented in Section 2 are provided in Section 5,

2. The truncated sample frequency spectrum.

2.1. *Background.* We denote Kingman's coalescent [27, 28, 29] on n leaves $\{\mathcal{C}_t^n\}_{t \geq 0}$ to be the backward-in-time Markov jump process, whose value at time t is a partition of $\{1, \dots, n\}$, and at time t , each pairs of blocks in \mathcal{C}_t^n coalesce with rate $\alpha(t)$. We also call $\frac{1}{\alpha(t)}$ the *population size history function*. We denote the ancestral process $A_t^{\mathcal{C}^n} = |\mathcal{C}_t^n|$ to be the number of blocks in \mathcal{C}_t^n , so that $A_t^{\mathcal{C}^n}$ is a pure death process with $A_0^{\mathcal{C}^n} = n$ and the rate of transition from m to $m - 1$ given by $\lambda_{m,m-1}^{\mathcal{C}}(t) = \binom{m}{2}\alpha(t)$.

We often drop the dependence on n , and write $\mathcal{C}_t = \mathcal{C}_t^n$ and $A_t^{\mathcal{C}} = A_t^{\mathcal{C}^n}$. We prefer to denote a dependence on n through the probability \mathbb{P}_n and the expectation \mathbb{E}_n . So if $X(\mathcal{C}^n)$ denotes a random variable of the process \mathcal{C}^n , we usually write $\mathbb{E}_n[X]$ instead of $\mathbb{E}[X(\mathcal{C}^n)]$.

Let ξ_i denote the partition of $\{1, \dots, n\}$ when \mathcal{C}_t has i blocks (also referred to as lineages). Let $T_i = \int_0^\infty \mathbb{I}_{A_t^{\mathcal{C}}=i} dt$ denote the amount of time \mathcal{C}_t has exactly i lineages. It is a fundamental fact of the coalescent that the waiting times $\mathbf{T}_{n:2} = (T_n, \dots, T_2)$ are independent of the partitions $\boldsymbol{\xi}_{n:2} = (\xi_n, \dots, \xi_2)$ [27].

A sample path of \mathcal{C}^n can be viewed as a rooted ultrametric tree with n leaves labeled $1, \dots, n$, where \mathcal{C}_t is the partition induced on $\{1, \dots, n\}$ by cutting the tree at height t . Now suppose we drop mutations onto this tree as a Poisson point process with rate $\frac{\theta}{2}$, and let \mathcal{M} denote the set of leaves that are beneath mutations (where we only consider mutations beneath the root, so by assumption $\mathcal{M} \neq \{1, \dots, n\}$). Then we define the sample frequency spectrum $f_n(k)$, for $0 < k < n$, as the first order Taylor series coefficient of $\mathbb{P}_n(|\mathcal{M}| = k)$ in the mutation rate,

$$\mathbb{P}_n(|\mathcal{M}| = k) = \frac{\theta}{2} f_n(k) + o(\theta).$$

We will generally refer to $f_n(k)$ as the sample frequency spectrum (SFS).

We also note two alternative definitions of the SFS. First, $f_n(k)$ is the expected number of mutations with k descendants when $\frac{\theta}{2} = 1$. Second,

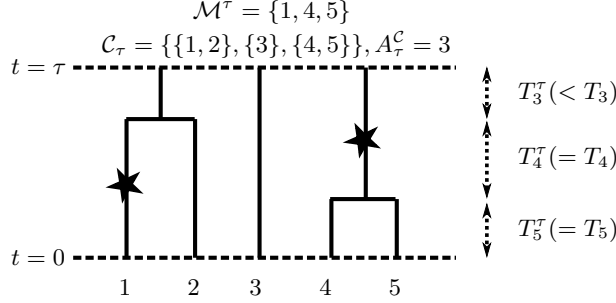


Fig 1: A sample path of the coalescent truncated at time τ . Star symbols denote mutations, while \mathcal{M}^τ denotes the set of leaves under those mutations. T_k^τ denotes the waiting time in the interval $[0, \tau)$ while there are k lineages.

$\frac{1}{\binom{n}{|K|}} f_n(|K|)$ is the expected length of the branch whose leaf set is K . More specifically, let \mathbb{I} denote the indicator function, and define $\mathcal{L}_K := \int_0^\infty \mathbb{I}_{K \in \mathcal{C}_t} dt$. Then

$$\frac{1}{\binom{n}{|K|}} f_n(|K|) = \mathbb{E}_n[\mathcal{L}_K].$$

The equivalence of these alternate definitions follows from previous results in Bhaskar, Kamm and Song [4], Griffiths and Tavaré [18], Jenkins and Song [23].

Note the SFS is sometimes defined to be a normalized version of $f_n(k)$, so that the entries sum to 1. We do not follow that convention, and use the unnormalized definition for the SFS throughout this paper.

2.2. The truncated coalescent and SFS. We now consider truncating the coalescent with mutation at time τ , as illustrated in Figure 1. Let \mathcal{M}^τ denote the set of leaves under mutations occurring in the time interval $[0, \tau)$. We define the *truncated* SFS $f_n^\tau(k)$ according to

$$\mathbb{P}_n(|\mathcal{M}^\tau| = k) = \frac{\theta}{2} f_n^\tau(k) + o(\theta).$$

By the same arguments as for the untruncated SFS, one can show that $f_n^\tau(k)$ gives the expected number of mutations in $[0, \tau)$ with k descendants, and letting $\mathcal{L}_K^\tau := \int_0^\tau \mathbb{I}_{K \in \mathcal{C}_t} dt$ denote the branch length subtending $K \subset \{1, \dots, n\}$ within $[0, \tau)$, we have

$$\frac{1}{\binom{n}{|K|}} f_n^\tau(|K|) = \mathbb{E}_n[\mathcal{L}_K^\tau].$$

Note that for $k < n$, we have $f_n(k) = f_n^\infty(k)$. For the truncated SFS, we will also consider mutations above the root, and so allow $k = n$ (i.e., $\mathcal{M}^\tau = \{1, \dots, n\}$), with $f_n^\tau(n) = \mathbb{E}_n[\mathcal{L}_{\{1, \dots, n\}}^\tau]$ giving the expected number of mutations within $[0, \tau)$ subtending the whole sample.

Given a random variable X , we define conditional versions of the SFS $f_n^\tau(k \mid X)$ according to

$$\mathbb{P}_n(|\mathcal{M}^\tau| = k \mid X) = \frac{\theta}{2} f_n^\tau(k \mid X) + o(\theta).$$

An example of a useful conditional SFS is $f_n^\tau(k \mid A_\tau^C = m)$, the expected branch length subtending k leaves given m ancestors at time τ . In particular, Chen [8] devised a dynamic program algorithm to compute the joint SFS for multiple populations under complex demographic histories, by computing $\{f_\nu^\tau(k)\}_{k \leq \nu \leq n}$ on each subpopulation of the history, where τ is the length of time a particular subpopulation exists. The unconditional SFS $f_\nu^\tau(k)$ is in turn computed in terms of $f_\nu^\tau(k \mid A_\tau^C = m)$ by writing

$$(1) \quad f_\nu^\tau(k) = \sum_{m=1}^{n-k+1} \mathbb{P}_\nu(A_\tau^C = m) f_\nu^\tau(k \mid A_\tau^C = m).$$

In Section 3.1, we describe the dynamic program algorithm for computing the joint SFS for multiple populations, and the way in which this algorithm uses the terms $f_\nu^\tau(k)$.

We consider how to compute (1). The first term in the summand, $\mathbb{P}_\nu(A_\tau^C = m)$, can be computed in at least three ways: by numerically exponentiating the rate matrix of A^C , by computing an alternating sum with $O(\nu)$ terms [41], or by solving a recursion we describe in Section 5.1. We note that the recursion described in Section 5.1 has the advantage of computing all values of $\mathbb{P}_\nu(A_\tau^C = m)$, $m \leq \nu \leq n$, in $O(n^2)$ time.

The second term $f_\nu^\tau(k \mid A_\tau^C = m)$ in the summand of (1) is computed in Chen [8] as

$$(2) \quad f_\nu^\tau(k \mid A_\tau^C = m) = \sum_{i=m}^{\nu} i p_{\nu,i}^{k,1} \mathbb{E}_\nu[T_i^\tau \mid A_\tau^C = m],$$

where

$$p_{\nu,i}^{k,j} := \begin{cases} \frac{\binom{k-1}{j-1} \binom{\nu-k-1}{i-j-1}}{\binom{\nu-1}{i-1}}, & \text{if } k \geq j > 0 \text{ and } \nu - k \geq i - j > 0, \\ 1, & \text{if } j = k = 0 \text{ or } i - j = \nu - k = 0, \\ 0, & \text{else,} \end{cases}$$

is the transition probability of the Pólya urn model, starting with $i - j$ white balls and j black balls, and ending with $\nu - k$ white balls and k black balls [24], and

$$T_i^\tau := \int_0^\tau \mathbb{I}_{A_t^C=i} dt$$

is the length of time in $[0, \tau)$ where there are i ancestral lineages to the sample, as illustrated in Figure 1. Chen [8] provides a formula for the conditional expectation $\mathbb{E}_\nu[T_i^\tau \mid A_\tau^C = m]$ for the case of constant population size, which he later extends [9] to the case of an exponentially growing population. However, these formulas involve a large alternating sum with $O(\nu^2)$ terms. Thus, computing $\mathbb{E}_\nu[T_i^\tau \mid A_\tau^C = m]$ for every value of i, m, ν , as required to compute $\{f_\nu^\tau(k)\}_{k \leq \nu \leq n}$ with (1) and (2), takes $O(n^5)$ time with these formulas. In addition, large alternating sums are numerically unstable due to catastrophic cancellation [20], and so these formulas require the use of high-precision numerical libraries, further increasing runtime.

2.3. An efficient, numerically stable algorithm for computing the truncated SFS. Here, we present a numerically stable algorithm to compute, for a given positive integer n , all of $\{f_\nu^\tau(k) \mid 1 \leq k \leq \nu \leq n\}$ in $O(n^2)$ time instead of $O(n^5)$ time. Our approach utilizes the following two lemmas:

LEMMA 1. *The entry $f_n^\tau(n)$ of the truncated SFS is given by*

$$(3) \quad f_n^\tau(n) = \tau - \sum_{k=1}^{n-1} \frac{k}{n} f_n^\tau(k).$$

LEMMA 2. *For all $1 \leq k \leq \nu$, the truncated SFS $f_\nu^\tau(k)$ satisfies the linear recurrence*

$$(4) \quad f_\nu^\tau(k) = \frac{\nu - k + 1}{\nu + 1} f_{\nu+1}^\tau(k) + \frac{k + 1}{\nu + 1} f_{\nu+1}^\tau(k + 1).$$

We prove Lemma 1 in Section 5.2. We note here that our proof also yields the identity $\mathbb{E}[T_{\text{MRCA}}] = \sum_{k=1}^{n-1} \frac{k}{n} f_n^\tau(k)$, where T_{MRCA} denotes the time to the most recent common ancestor of the sample; to our knowledge, this formula is new. A proof of Lemma 2 is provided in Section 5.3.

We now sketch our algorithm. For a given n , we show below that all values of $f_n^\tau(k)$, for $1 \leq k < n$, can be computed in $O(n^2)$ time. We then compute $f_n^\tau(n)$ using Lemma 1 in $O(n)$ time. Finally, using $f_n^\tau(k)$ for $1 \leq k \leq n$ as boundary conditions, Lemma 2 allows us to compute all $f_\nu^\tau(k)$, for $\nu = n - 1, n - 2, \dots, 2$ and $k = 1, \dots, \nu$, in $O(n^2)$ time.

We now describe how to compute the aforementioned terms $f_n^\tau(k)$, for all $k < n$, in $O(n^2)$ time. We first recall the result of Polanski and Kimmel [37] which represents the untruncated SFS $f_n(k)$, for $1 \leq k \leq n-1$, as

$$(5) \quad f_n(k) = \sum_{m=2}^n W_{n,k,m} c_m,$$

where

$$(6) \quad \begin{aligned} c_m &:= \mathbb{E}_m[T_m] = \int_0^\infty t \binom{m}{2} \alpha(t) \exp \left[- \binom{m}{2} \int_0^t \alpha(x) dx \right] dt \\ &= \int_0^\infty \exp \left[- \binom{m}{2} \int_0^t \alpha(x) dx \right] dt \end{aligned}$$

denotes the waiting time to the first coalescence for a sample of size m , and $W_{n,k,m}$ are universal constants that are efficiently computable using the following recursions [37]:

$$(7) \quad \begin{aligned} W_{n,k,2} &= \frac{6}{n+1}, \\ W_{n,k,3} &= 30 \frac{(n-2k)}{(n+1)(n+2)}, \\ W_{n,k,m+2} &= -\frac{(1+m)(3+2m)(n-m)}{m(2m-1)(n+m+1)} W_{n,k,m} + \frac{(3+2m)(n-2k)}{m(n+m+1)} W_{n,k,m+1}, \end{aligned}$$

for $2 \leq m \leq n-2$. The key observation is to note that, in a similar vein as (5), we have:

LEMMA 3. *The truncated SFS $f_n^\tau(k)$, for $1 \leq k \leq n-1$, can be written as*

$$(8) \quad f_n^\tau(k) = \sum_{m=2}^n W_{n,k,m} c_m^\tau,$$

where c_m^τ is a truncated version of (6):

$$(9) \quad c_m^\tau := \mathbb{E}_m[T_m^\tau] = \int_0^\tau \exp \left[- \binom{m}{2} \int_0^t \alpha(x) dx \right] dt.$$

We prove Lemma 3 in Section 5.4. For piecewise-exponential $\alpha(t)$, c_m^τ can be computed explicitly using formulas from Bhaskar, Wang and Song [5]. Using (7), we can compute all values of $W_{n,k,m}$, for $1 \leq k \leq n$ and $2 \leq m \leq n$, in $O(n^2)$ time. Then, using (8), all values of $f_n^\tau(k)$, for $1 \leq k \leq n-1$ can be computed in $O(n^2)$ time.

Note that the above algorithm not only significantly improves computational complexity, but also resolves numerical issues, since it allows us to avoid computing the expected times $\mathbb{E}_\nu[T_i^\tau \mid A_\tau^C = m]$, which are alternating sums of $O(n^2)$ terms and are numerically unstable to evaluate for large values of n (say, $n > 50$).

2.4. An alternative formula for piecewise-constant subpopulation sizes.

For demographic scenarios with piecewise-constant subpopulation sizes, we present an alternative formula for computing the truncated SFS within a constant piece. This formula has the same sample computational complexity as that described in the previous section.

Let \mathcal{K}_t denote the *coalescent with killing*, a stochastic process that is closely related to the Chinese restaurant process, Hoppe's urn, and Ewens' sampling formula [2, 21]. In particular, the coalescent with killing $\{\mathcal{K}_t\}_{t \geq 0}$ is a stochastic process whose value at time t is a *marked* partition of $\{1, \dots, n\}$, where each partition block is marked as “killed” or “unkilled”. We obtain the partition for \mathcal{K}_t by dropping mutations onto the coalescent tree as a Poisson point process with rate $\frac{\theta}{2}$, and then defining an equivalence relation on $\{1, \dots, n\}$, where $i \sim j$ if and only if i, j have coalesced by time t and there are no mutations on the branches between i and j (i.e., i and j are identical by descent). We furthermore mark the equivalence classes (i.e. partition blocks) of \mathcal{K}_t that are descended from a mutation in $[0, t)$ as “killed”. See Figure 2 for an illustration. The process \mathcal{K}_τ can also be obtained by running Hoppe's urn, or equivalently the Chinese restaurant process, forward in time [12, Theorem 1.9].

Let A_t^K be the number of unkilld blocks in \mathcal{K}_t , so that A_t^K is a pure death process with transition rate $\lambda_{i,i-1}^K(t) = \binom{i}{2}\alpha(t) + \frac{i\theta}{2}$ (the rate of coalescence is the number of unkilld pairs $\binom{i}{2}\alpha(t)$, and the rate of killing due to mutation is $\frac{i\theta}{2}$). Our next proposition gives a formula for the truncated conditional sample frequency spectrum given A_τ^K , i.e., $f_n^\tau(k \mid A_\tau^K = m)$.

PROPOSITION 1. *Consider the constant population size history $\frac{1}{\alpha(t)} = \frac{1}{\alpha}$ for $t \in [0, \tau)$, and let $m > 0$ and $0 < k \leq n - m$. The joint probability that the number of derived mutants is k and the number of unkilld ancestral*

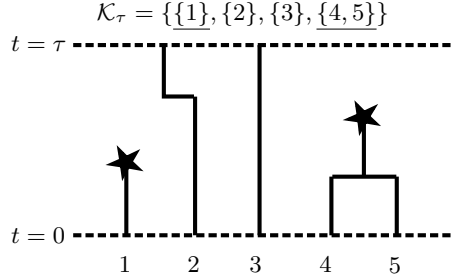


Fig 2: The coalescent with killing for the genealogy in Figure 1. Note that \mathcal{K}_τ is a marked partition, with the blocks killed by mutations in $[0, \tau)$ being specially marked.

lineages is m , when truncating at time τ , is given by

$$\mathbb{P}_n(|\mathcal{M}^\tau| = k, A_\tau^\mathcal{K} = m) = \frac{\theta}{2} f_n^\tau(k \mid A_\tau^\mathcal{K} = m) \mathbb{P}(A_\tau^\mathcal{C} = m) + o(\theta),$$

where

$$(10) \quad f_n^\tau(k \mid A_\tau^\mathcal{K} = m) = \frac{2}{\alpha k} \frac{\binom{n-m}{k}}{\binom{n-1}{k}}.$$

We prove Proposition 1 in Section 5.5. Note that this equation does not hold for the case $k = n, m = 0$, but fortunately we do not need to consider that case in what follows below.

We can use Proposition 1 to stably and efficiently compute the terms $f_n^\tau(k)$, for $k \leq \nu \leq n$, as follows. We first compute the case $k < \nu = n$. Note that $\mathbb{P}_n(|\mathcal{M}^\tau| = K) = \sum_m \mathbb{P}_n(|\mathcal{M}^\tau| = K, A_\tau^\mathcal{K} = m)$. So for $k < n$, by Proposition 1

$$(11) \quad \begin{aligned} f_n^\tau(k) &= \sum_{m=1}^n f_n^\tau(k \mid A_\tau^\mathcal{K} = m) \mathbb{P}_n(A_\tau^\mathcal{C} = m) \\ &= \sum_{m=1}^n \frac{2}{\alpha k} \frac{\binom{n-m}{k}}{\binom{n-1}{k}} \mathbb{P}_n(A_\tau^\mathcal{C} = m). \end{aligned}$$

The sum in (11) contains $O(n)$ terms, so it costs $O(n^2)$ to compute $f_n^\tau(k)$ for all $k < n$. After this, we use Lemma 1 to compute $f_n^\tau(n)$, and then use Lemma 2 to compute $f_n^\tau(k)$ for all $1 \leq k \leq \nu < n$. Since there are $O(n^2)$ such terms, this also takes $O(n^2)$ time.

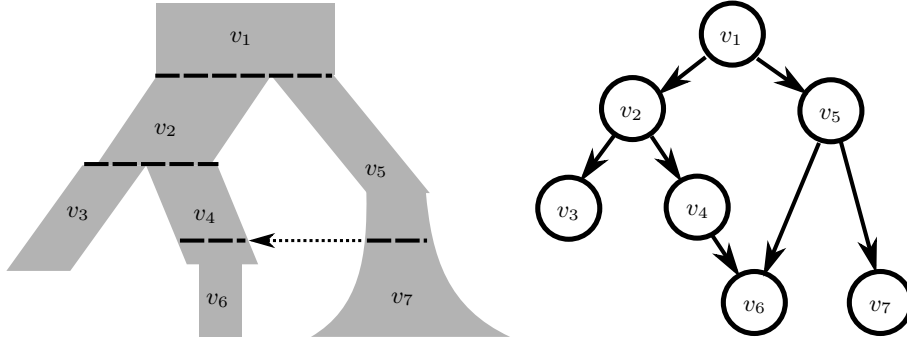


Fig 3: A demographic history with a pulse migration event (left), and its corresponding directed graph (right).

3. The joint SFS for multiple populations. In this section we discuss an algorithm for computing the multi-population SFS [8, 9, 42]. We describe the algorithm in Section 3.1, and note how the results from Section 2 improve the time complexity of this algorithm. In Section 3.2, we focus on the special case of tree-shaped demographies, and introduce a further algorithmic speedup by replacing the coalescent with a Moran model.

Let V be the number of subpopulations in the demographic history, n the total sample size, and L the number of SFS entries to compute. Then the results from Section 2 improve the computational complexity of the SFS from $O(n^5V + WL)$ to $O(n^2V + WL)$, where W is a term that depends on the structure of the demographic history. In the special case of tree-shaped demographies, the algorithm from Chen [8] gives $W = O(n^4V)$. The Moran-based speedup from Section 3.2, combined with results from Bryant et al. [7], improves this to $W = O(n \log(n)V)$.

The Moran-based speedup can be generalized to non-tree demographies, but the notation, implementation, and analysis of computational complexity becomes substantially more complicated. We thus leave its generalization to future work.

3.1. A coalescent-based dynamic program. Suppose at the present we have \mathcal{D} populations, and in the i th population we observe n_i alleles. For a single point mutation, let $\mathbf{x} = (x_1, \dots, x_{\mathcal{D}})$ denote the number of alleles that are derived in each population. We wish to compute $f(\mathbf{x})$, where $\frac{\theta}{2}f(\mathbf{x})$ is the expected number of point mutations with derived counts \mathbf{x} .

For demographic histories consisting of population size changes, population splits, population mergers, and pulse admixture events, Chen [8] gave an algorithm to compute $f(\mathbf{x})$ using the truncated SFS $f_n^\tau(k)$ that we defined

in Section 2.

We describe this algorithm to compute $f(\mathbf{x})$. We start by representing the population history as a directed acyclic graph (DAG), where each vertex v represents a subpopulation (Figure 3). We draw a directed edge from v to v' if there is gene flow from the bottom-most part of v to the top-most part of v' , where “down” is the present and “up” is the ancient past. Thus, the leaf vertices correspond to the subpopulations at the present. For a vertex v in the population history graph, let $\tau_v \in (0, \infty)$ denote the length of time the corresponding population persists, and let $\alpha_v : [0, \tau_v) \rightarrow \mathbb{R}^+$ denote the inverse population size history of v . So going backwards in time from the present, $\alpha_v(t)$ gives the instantaneous rate at which two lineages in v coalesce, after v has existed for time t . We use $f_n^v(k)$ to denote the truncated SFS for the coalescent embedded in v , i.e., $f_n^v(k) = f_n^{\tau_v}(k)$ for a coalescent with coalescence rate $\alpha_v(t)$. Then we have

$$(12) \quad f(\mathbf{x}) = \sum_v \sum_{m_0^v, k_0^v} f_{m_0^v}^v(k_0^v) \mathbb{P}(\mathbf{x} \mid k_0^v, m_0^v) \mathbb{P}(m_0^v)$$

where m_0^v denotes the number of lineages at the bottom of v that are ancestral to the initial sample, and k_0^v denotes the number of these lineages with a derived allele.

In order to use (12), we must compute $f_{m_0^v}^v(k_0^v)$ for every population v , and every value of m_0^v and k_0^v . If n is the total sample size and V the total number of vertices, then this takes $O(n^5 V)$ time using the formulas of Chen [8]. Our results from Section 2 improve this to $O(n^2 V)$.

To use (12), we must also compute the terms $\mathbb{P}(\mathbf{x} \mid k_0^v, m_0^v) \mathbb{P}(m_0^v)$, for which Chen [8] constructs a dynamic program, starting at the leaf vertices and moving up the graph. This dynamic program essentially consists of setting up a Bayesian graphical model with random variables m_0^v, k_0^v and performing belief propagation, which can be done via the sum-product algorithm (“tree-peeling”) if the population graph is a tree [14, 35], or via a junction tree algorithm if not [30].

The time complexity of the algorithm thus depends on the topological structure of the population graph. In the special case where the demographic history is a binary tree, the tree-peeling algorithm computes the values $\mathbb{P}(\mathbf{x} \mid k_0^v, m_0^v) \mathbb{P}(m_0^v)$ in $O(n^4 V)$ time, since the vertex v has $O(n^2)$ possible states (k_0^v, m_0^v) , so summing over the transitions between every pair of states costs $O(n^4)$. Note that Chen [8] mistakenly states that the computation takes $O(n^3 V)$ time. In the further special case that the population sizes are piecewise constant, speedups from Bryant et al. [7] can improve this to $O(n^2 \log(n) V)$. More specifically, Bryant et al. [7] computes the terms

$\mathbb{P}(\mathbf{x} \mid k_0^v, m_0^v) \mathbb{P}(m_0^v)$ in $O(n^2 \log(n)V)$ time for a model with recurrent mutation, but the results can be applied straightforwardly here by setting the mutation rate to 0, thus disallowing recurrent mutation.

To summarize, let W be the time it takes to compute (12) after the terms $f_m^v(k)$ have been precomputed, and let L be the number of distinct entries \mathbf{x} for which we wish to compute $f(\mathbf{x})$. Then our results from Section 2 improve the computational complexity from $O(n^5V + WL)$ to $O(n^2V + WL)$. In the case of a binary tree the original algorithm of Chen [8] gives $W = O(n^4V)$, but adapting results from Bryant et al. [7] improves this to $W = O(n^2 \log(n)V)$ when the population sizes are piecewise constant. In the following section, we introduce a new approach that further improves the runtime to $W = O(n \log(n)V)$ and generalizes from piecewise constant to arbitrary population size histories.

3.2. A Moran-based dynamic program. We describe a modified version of the dynamic programs from Bryant et al. [7], Chen [8] that improves the computational complexity of computing $f(\mathbf{x})$ for tree-shaped demographies. The main idea is to replace the backwards-in-time coalescent with a forwards-in-time Moran model.

We assume the \mathcal{D} populations at the present are related by a binary rooted tree with \mathcal{D} leaves, where each leaf represents a population at the present, and at each internal vertex, a parent population splits into two child populations. (Note that a non-binary tree can be represented as a binary tree, with additional vertices of height 0).

Instead of working with the multi-population coalescent directly, we will consider a multi-population Moran model, in which the coalescent is embedded [32]. In particular, let $\mathcal{L}(v)$ denote the leaf populations descended from the population v , and let $n_v = \sum_{i \in \mathcal{L}(v)} n_i$ be the number of present-day alleles with ancestry in v . For each population v (except the root), we construct a Moran model going *forward* in time, i.e. starting at τ_v and ending at 0. The Moran model consists of n_v lineages, each with either an ancestral or derived allele. Going forward in time, every lineage copies itself onto every other lineage at rate $\frac{1}{2}\alpha_v(t)$. Thus, the total rate of copying events is $\binom{n_v}{2}\alpha_v(t)$. Let μ_t^v denote the number of derived alleles at time t in population v . Then the transition rate of μ_t^v when $\mu_t^v = x$ is $\lambda_{x \rightarrow x+1}(t) = \lambda_{x \rightarrow x-1}(t) = \frac{x(n_v - x)}{2}\alpha_v(t)$, since there are $x(n_v - x)$ pairs of lineages with different alleles.

The coalescent is embedded within the Moran model, because if we trace the ancestry of genetic material backwards in time in the Moran model, we obtain a genealogy with the same distribution as under the coalescent (Theorem 1.30 of Durrett [12]). Thus, we can obtain the expected number

of mutations with derived counts \mathbf{x} , by summing over which population v the mutation occurred in:

$$(13) \quad f(\mathbf{x}) = \sum_v \sum_{k=1}^{n_v} f_{n_v}^v(k) \mathbb{P}(\mathbf{x} \mid \mu_0^v = k).$$

Let $\mathbf{x}_v = \{x_i : i \in \mathfrak{L}(v)\}$ denote the subsample of derived allele counts in the populations descended from v . Similarly, let $\mathbf{x}_v^c = \{x_i : i \notin \mathfrak{L}(v)\}$. Then for $k \geq 1$,

$$(14) \quad \mathbb{P}(\mathbf{x} \mid \mu_0^v = k) = \begin{cases} \mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k), & \text{if } \mathbf{x}_v^c = \mathbf{0}, \\ 0, & \text{if } \mathbf{x}_v^c \neq \mathbf{0}. \end{cases}$$

So it suffices to compute $\mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k)$ for all v and k . If v is the i th leaf population, then $\mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k) = \mathbb{I}_{k=x_i}$. On the other hand, if v is an interior vertex with children v_1 and v_2 , then

$$(15) \quad \mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k) = \sum_{k_1=0}^{n_{v_1}} \frac{\binom{n_{v_1}}{k_1} \binom{n_{v_2}}{k-k_1}}{\binom{n_v}{k}} \mathbb{P}(\mathbf{x}_{v_1} \mid \mu_{\tau_{v_1}}^{v_1} = k_1) \mathbb{P}(\mathbf{x}_{v_2} \mid \mu_{\tau_{v_2}}^{v_2} = k - k_1),$$

where $\mathbb{P}(\mathbf{x}_{v_i} \mid \mu_{\tau_{v_i}}^{v_i})$ can be computed from

$$(16) \quad \mathbb{P}(\mathbf{x}_v \mid \mu_{\tau_v}^v = k) = \sum_{j=0}^{n_v} \mathbb{P}(\mathbf{x}_v \mid \mu_0^v = j) \mathbb{P}(\mu_0^v = j \mid \mu_{\tau_v}^v = k).$$

To compute the transition probability $\mathbb{P}(\mu_0^v = j \mid \mu_{\tau_v}^v = k)$, note that the transition rate matrix of μ_t^v can be written as $Q^{(v)}\alpha(t)$, where $Q^{(v)} = (q_{ij}^{(v)})_{0 \leq i, j \leq n_v}$ is a $(n+1) \times (n+1)$ matrix with

$$q_{ij}^{(v)} = \begin{cases} -i(n_v - i), & \text{if } i = j, \\ \frac{1}{2}i(n_v - i), & \text{if } |j - i| = 1, \\ 0, & \text{else,} \end{cases}$$

so then the transition probability is given by the matrix exponential

$$(17) \quad \mathbb{P}(\mu_0^v = j \mid \mu_{\tau_v}^v = k) = (e^{Q^{(v)} \int_0^{\tau_v} \alpha_v(t) dt})_{k,j}.$$

Thus, the joint SFS $f(\mathbf{x})$ can be computed using (13) and (14), with $\mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k)$ given by recursively computing (15), (16), and (17), in a depth-first search on the population tree (i.e. Felsenstein's tree-peeling algorithm, or the sum-product algorithm for belief propagation).

We now consider the computational complexity associated with each vertex v . Equations (15) and (16) each have $O(n_v)$ terms, and must be solved for $O(n_v)$ values of k ; so naively, each vertex costs $O(n_v^2)$ time. However, we can improve (15) to $O(n_v \log(n_v))$ and (16) to $O(n_v)$, using essentially the same speedups as in Bryant et al. [7]. Letting $\tilde{\ell}_t^v(k) = \binom{n_v}{k} \mathbb{P}(\mathbf{x}_v \mid \mu_t^v = k)$, (15) can be written as a convolution

$$(18) \quad \tilde{\ell}_0^v = \tilde{\ell}_{\tau_{v1}}^{v1} * \tilde{\ell}_{\tau_{v2}}^{v2},$$

which can be computed in $O(n_v \log(n_v))$ time via the fast Fourier transform [10], since $\mathcal{F}\ell_0^v = \left(\mathcal{F}\ell_{\tau_{v1}}^{v1}\right) \left(\mathcal{F}\ell_{\tau_{v2}}^{v2}\right)$, where \mathcal{F} is the discrete Fourier transform. Similarly, letting $\ell_t^v(k) = \tilde{\ell}_t^v(k) / \binom{n_v}{k}$, (16) turns into

$$(19) \quad \ell_{\tau_v}^v = e^{(Q^{(v)} \int_0^{\tau_v} \alpha_v(t) dt)} \ell_0^v,$$

and this costs $O(n_v)$ by the sparsity of $Q^{(v)}$, using results for computing the action of sparse matrix exponentials [1, 39]. Transforming between $\tilde{\ell}_{\tau_v}^v$ and $\ell_{\tau_v}^v$ takes $O(n_v)$ time.

The computational complexity associated with a single vertex v is thus $O(n_v \log(n_v))$. Therefore, computing the joint SFS entry $f(\mathbf{x})$ for L distinct values of \mathbf{x} takes $O(n^2 V + n \log(n) V L)$ time for a binary population tree with arbitrary population size functions and no migration. This is a substantial improvement over the $O(n^5 V + n^4 V L)$ complexity of Chen [8], and the $O(n^2 \log(n) V L)$ complexity of Bryant et al. [7]. Similar to Chen [8], our approach has the benefit of easily generalizing to arbitrary population size histories, not just piecewise constant sizes.

4. Results. We implemented our formulas and algorithm in Python, using the Python packages *numpy* and *scipy*. We also implemented the formulas from Chen [8, 9], and compared the performance of the two algorithms on simulated data.

We simulated data for demographic trees with $\mathcal{D} \in \{5, 10, 15, 25, 50, 100\}$ populations at the present, and $\frac{n}{\mathcal{D}} \in \{1, 2, 5, 10\}$ individuals per population. For each value of n, \mathcal{D} , we used the program *scrm* [40] to generate 20 random datasets, each with a demographic history that is a random binary tree.

In Figure 4, we compare the running time of the original algorithm of Chen [8, 9] against our new algorithm that utilizes the formulas for $f_n^\tau(k)$ presented in Section 2 and our new Moran-based approach described in Section 3.2. We find our algorithm to be orders of magnitude faster; the difference is especially pronounced as the number \mathcal{D} of populations grows.

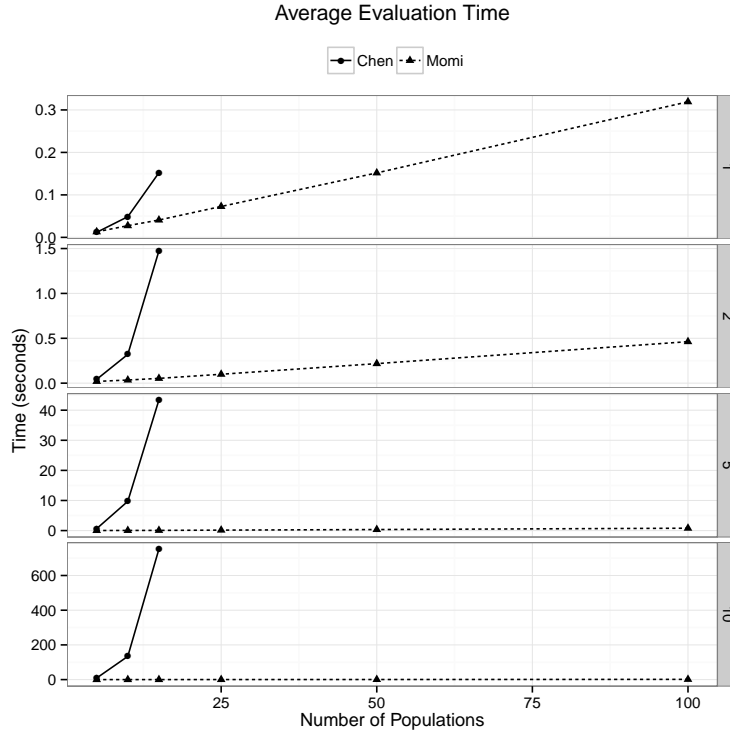


Fig 4: Average computation time per joint SFS entry. For each combination of the number \mathcal{D} of populations and the sample size n/\mathcal{D} per population, we generated 20 random datasets, each under a demographic history that is a random binary tree. The expected joint SFS for the resulting segregating sites were then computed using our method (*mom*) and that of Chen [8]. Average runtime (in seconds) per joint SFS entry is plotted on the y -axis, with each panel corresponding to a different value of n/\mathcal{D} . As the plots show, our algorithm is orders of magnitude faster than Chen’s. Due to its significantly increased runtime, we were able to run Chen’s method only up to $\mathcal{D} = 15$.

Note that, due to the increased running time, we were only able to run Chen’s algorithm to completion for $\mathcal{D} \leq 15$.

In Figure 5, we compare the accuracy of the two algorithms. The figure compares the SFS entries returned by the two methods across a subset of the simulations depicted in Figure 4. To adequately capture the large range of numerical values returned by the Chen method, we transformed each SFS

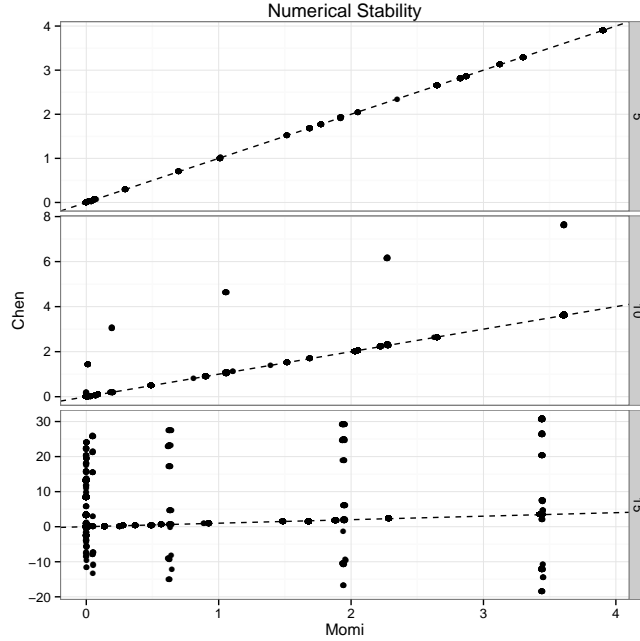


Fig 5: Numerical stability of the two algorithms. The plot compares the numerical values returned by our method (*momi*) and Chen’s method, for the simulations described in Figure 4. The three panels on the y -axis correspond to $\mathcal{D} \in \{5, 10, 15\}$. To adequately illustrate the observed range of numerical values, the SFS values were transformed via the map $z \mapsto \text{sign}(z) \log_{10}(1 + |z|)$; the dashed line represents the identity $y = x$. The two methods agree for $\mathcal{D} \leq 5$, but Chen’s method displays considerable numerical instability for $\mathcal{D} \geq 10$.

entry using the transformation $z \mapsto \text{sign}(z) \log_{10}(1 + |z|)$. The line $y = x$ is also plotted; points falling on the line depict the SFS entries where both methods agreed. All negative return values represent numerical errors. The two methods agree for $\mathcal{D} \leq 5$, but Chen’s algorithm displays considerable numerical instability for $\mathcal{D} = 10$ and higher.

5. Proofs. In this section, we provide proofs of the mathematical results presented in earlier sections.

5.1. *A recursion for efficiently computing $\mathbb{P}_\nu(A_\tau^C = m)$.* We describe how to compute $\mathbb{P}_\nu(A_\tau^C = m)$, for all values of $m \leq \nu \leq n$, in $O(n^2)$ time. First,

note that

$$\begin{aligned}
\mathbb{P}_{\nu-1}(A_\tau^{\mathcal{C}} = m) &= \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1, \{\nu\} \in \mathcal{C}_\tau) + \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m, \{\nu\} \notin \mathcal{C}_\tau) \\
&= \frac{(m+1)p_{\nu,m+1}^{1,1}}{\binom{\nu}{1}} \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1) + \left(1 - \frac{mp_{\nu,m}^{1,1}}{\binom{\nu}{1}}\right) \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m) \\
&= \frac{(m+1)(m)}{\nu(\nu-1)} \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1) + \left(1 - \frac{m(m-1)}{\nu(\nu-1)}\right) \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m).
\end{aligned}$$

Rearranging, we get the recursion

(20)

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m) = \frac{1}{1 - \frac{m(m-1)}{\nu(\nu-1)}} \left[\mathbb{P}_{\nu-1}(A_\tau^{\mathcal{C}} = m) - \frac{(m+1)(m)}{\nu(\nu-1)} \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1) \right]$$

with base cases

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = \nu) = e^{-\binom{\nu}{2} \int_0^\tau \alpha(t) dt}.$$

So after solving $\int_0^\tau \alpha(t) dt$, we can use the recursion and memoization to solve for all of the $O(n^2)$ terms $\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m)$ in $O(n^2)$ time. In particular, in the case of constant population size, $\alpha(t) = \alpha$, the base case is given by

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = \nu) = e^{-\binom{\nu}{2} \alpha \tau},$$

and in the case of an exponentially growing population size, $\alpha(t) = \alpha(\tau)e^{\beta(\tau-t)}$, the base case is given by

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = \nu) = e^{-\binom{\nu}{2} \alpha(\tau)(e^{\beta\tau} - \frac{1}{\beta})}.$$

5.2. Proof of Lemma 1. Let T_{MRCA} denote the time to the most recent common ancestor of the sample. We first note that

$$f_n^\tau(n) = \tau - \mathbb{E}_n[T_{\text{MRCA}} \wedge \tau],$$

since the branch length subtending the whole sample is the time between τ and T_{MRCA} .

Next, note that $\frac{\theta}{2} \mathbb{E}_n[T_{\text{MRCA}} \wedge \tau]$ is equal to the number of polymorphic mutations in $[0, \tau)$ where the individual “1” is derived. This is because, as we trace the ancestry of “1” backwards in time, all mutations hitting the lineage below T_{MRCA} are polymorphic, while all mutations hitting above T_{MRCA} are monomorphic.

The expected number of polymorphic mutations with “1” derived is also equal to $\frac{\theta}{2} \sum_{k=1}^{n-1} \frac{k}{n} f_n^\tau(k)$, since if a mutation has k derived leaves, the chance that “1” is in the derived set is $\frac{k}{n}$. Thus,

$$\mathbb{E}_n[T_{\text{MRCA}} \wedge \tau] = \sum_{k=1}^{n-1} \frac{k}{n} f_n^\tau(k),$$

which completes the proof.

5.3. *Proof of Lemma 2.* We first note that

$$\begin{aligned} \mathbb{P}_n(\mathcal{M}^\tau = \{1, \dots, k\}) \\ = \mathbb{P}_{n+1}(\mathcal{M}^\tau = \{1, \dots, k\}) + \mathbb{P}_{n+1}(\mathcal{M}^\tau = \{1, \dots, k, n+1\}). \end{aligned}$$

By exchangeability, we have $\mathbb{P}_n(\mathcal{M}^\tau = K) = \frac{\theta}{2} \frac{f_n^\tau(|K|)}{\binom{n}{|K|}} + o(\theta)$ for all $K \subseteq \{1, \dots, n\}$, so

$$\frac{1}{\binom{n}{k}} f_n^\tau(k) = \frac{1}{\binom{n+1}{k}} f_{n+1}^\tau(k) + \frac{1}{\binom{n+1}{k+1}} f_{n+1}^\tau(k+1).$$

Multiplying both sides by $\binom{n}{k}$ gives

$$f_n^\tau(k) = \frac{n-k+1}{n+1} f_{n+1}^\tau(k) + \frac{k+1}{n+1} f_{n+1}^\tau(k+1).$$

5.4. *Proof of Lemma 3.* Let $\alpha^*(t)$ denote the inverse population size history given by

$$\alpha^*(t) = \begin{cases} \alpha(t) & \text{if } t < \tau \\ \infty & \text{if } t \geq \tau. \end{cases}$$

So the demographic history with population size $\frac{1}{\alpha^*(t)}$ agrees with the original history up to time τ , at which point the population size drops to 0, and all lineages instantly coalesce into a single lineage with probability 1.

Let $T_{m,*}$ denote the amount of time there are m ancestral lineages for the coalescent with size history $\frac{1}{\alpha^*(t)}$. Similarly, let $f_{n,*}(k)$ denote the SFS under the size history $\frac{1}{\alpha^*(t)}$. Then from the result of Polanski and Kimmel [37],

$$f_{n,*}(k) = \sum_{m=2}^n W_{n,k,m} \mathbb{E}_m[T_{m,*}].$$

Note that for $m > 1$, we almost surely have $T_{m,*} = T_{m,*}^\tau$, i.e. the intercoalescence time equals its truncated version, since all lineages coalesce instantly at τ with probability 1. Thus, $\mathbb{E}_m[T_{m,*}] = \mathbb{E}_m[T_{m,*}^\tau]$. Similarly, for $k < n$, $f_{n,*}(k) = f_{n,*}^\tau(k)$, i.e. the SFS equals the truncated SFS, because the probability of a polymorphic mutation occurring in $[\tau, \infty)$ is 0.

Finally, note that $\mathbb{E}_m[T_{m,*}^\tau] = \mathbb{E}_m[T_m^\tau]$ and $f_{n,*}^\tau(k) = f_n^\tau(k)$, because $\alpha(t)$ and $\alpha^*(t)$ are identical on $[0, \tau)$.

5.5. Proof of Proposition 1. We start by showing that $\mathbb{P}_n(A_\tau^\mathcal{K} = m) = \mathbb{P}_n(A_\tau^\mathcal{C} = m) + O(\theta)$. Let $T_i^\tau(\mathcal{K}) = \int_0^\tau \mathbb{I}_{A_t^\mathcal{K}=i} dt$ denote the amount of time where \mathcal{K} has i unkilld lineages. Let p denote the probability density function. For (t_n, \dots, t_m) with $\sum t_i = \tau$, we have

$$\begin{aligned} p(T_n^\tau(\mathcal{K}) = t_n, \dots, T_m^\tau(\mathcal{K}) = t_m) \\ &= e^{-\lambda_{m,m-1}^\mathcal{K} t_m} \prod_{i=m+1}^n \lambda_{i,i-1}^\mathcal{K} e^{-\lambda_{i,i-1}^\mathcal{K} t_i} \\ &= e^{-((\binom{m}{2}\alpha + \frac{m\theta}{2})t_m)} \prod_{i=m+1}^n \left(\binom{i}{2}\alpha + \frac{i\theta}{2} \right) e^{-((\binom{i}{2}\alpha + \frac{i\theta}{2})t_i)} \\ &= e^{-\binom{m}{2}\alpha t_m} \prod_{i=m+1}^n \binom{i}{2}\alpha e^{-\binom{i}{2}\alpha t_i} + O(\theta) \\ &= p(T_n^\tau = t_n, \dots, T_m^\tau = t_m) + O(\theta), \end{aligned}$$

and so

$$\begin{aligned} \lim_{\theta \rightarrow 0} \mathbb{P}_n(A_\tau^\mathcal{K} = m) &= \lim_{\theta \rightarrow 0} \int_{\sum t_i = \tau} p(T_n^\tau(\mathcal{K}) = t_n, \dots, T_m^\tau(\mathcal{K}) = t_m) d\mathbf{t} \\ &= \int_{\sum t_i = \tau} p(T_n^\tau = t_n, \dots, T_m^\tau = t_m) d\mathbf{t} \\ &= \mathbb{P}_n(A_\tau^\mathcal{C} = m). \end{aligned}$$

where we can exchange the limit and the integral by the Bounded Convergence Theorem, because $p(T_n^\tau(\mathcal{K}) = t_n, \dots, T_m^\tau(\mathcal{K}) = t_m) \leq \prod_{i=m+1}^n \left(\binom{i}{2}\alpha + \frac{i}{2} \right)$ for $\theta \leq 1$.

Thus we have

$$\begin{aligned} \mathbb{P}_n(|\mathcal{M}^\tau| = k, A_\tau^\mathcal{K} = m) &= \mathbb{P}_n(|\mathcal{M}^\tau| = k \mid A_\tau^\mathcal{K} = m) \mathbb{P}_n(A_\tau^\mathcal{K} = m) \\ &= \left(\frac{\theta}{2} f_n^\tau(k \mid A_\tau^\mathcal{K} = m) + o(\theta) \right) (\mathbb{P}_n(A_\tau^\mathcal{C} = m) + O(\theta)) \\ &= \frac{\theta}{2} f_n^\tau(k \mid A_\tau^\mathcal{K} = m) \mathbb{P}_n(A_\tau^\mathcal{C} = m) + o(\theta), \end{aligned}$$

which proves the first part of the proposition.

We next solve for $f_n^\tau(k \mid A_\tau^\mathcal{K} = m)$, the first order Taylor series coefficient for $\mathbb{P}_n(|\mathcal{M}^\tau| = k \mid A_\tau^\mathcal{K} = m)$ in the mutation rate $\frac{\theta}{2}$.

When there are i unkilld lineages, the probability that the next event is a killing event is $\frac{\theta}{\alpha(i-1)+\theta} = \frac{\theta}{\alpha(i-1)} + o(\theta)$. Given that the event is a killing, the chance that the killed lineage has k leaf descendants is $p_{n,i}^{k,1}$. So summing over i , and dividing out the mutation rate $\frac{\theta}{2}$, we get

$$\begin{aligned}
f_n^\tau(k \mid A_\tau^\mathcal{K} = m) &= \frac{2}{\alpha} \sum_{i=m+1}^{n-k+1} \frac{1}{i-1} p_{n,i}^{k,1} \\
&= \frac{2}{\alpha} \sum_{i=m+1}^{n-k+1} \frac{1}{i-1} \frac{\binom{n-k-1}{i-2}}{\binom{n-1}{i-1}} \\
&= \frac{2}{\alpha} \sum_{i=m+1}^{n-k+1} \frac{1}{i-1} \frac{(n-k-1)!(i-1)!(n-i)!}{(i-2)!(n-k-i+1)!(n-1)!} \\
&= \frac{2(n-k-1)!}{\alpha(n-1)!} \sum_{i=m+1}^{n-k+1} \frac{(n-i)!}{(n-k-i+1)!} \\
&= \frac{2(n-k-1)!}{\alpha(n-1)!} \sum_{j=0}^{n-k-m} \frac{(j+k-1)!}{j!} \\
&= \frac{2}{\alpha k \binom{n-1}{k}} \sum_{j=0}^{n-k-m} \binom{j+k-1}{j} \\
&= \frac{2}{\alpha k} \frac{\binom{n-m}{k}}{\binom{n-1}{k}},
\end{aligned}$$

where we made the change of variables $j = n - k - i + 1$, and where the final line follows from repeated application of the combinatorial identity $\binom{a}{b} = \binom{a-1}{b} + \binom{a-1}{b-1}$.

5.5.1. Alternative proof for $f_n^\tau(k \mid A_\tau^\mathcal{K} = m)$ via the Chinese Restaurant Process. We sketch an alternative proof of the expression for $f_n^\tau(k \mid A_\tau^\mathcal{K} = m)$, using the Chinese Restaurant Process.

Consider the coalescent with killing going forward in time (towards the present), and only looking at it when the number of individuals increases. Then when there are i lineages, a new mutation occurs with probability $\frac{\theta}{\alpha i + \theta} = \frac{\theta/\alpha}{i + \theta/\alpha}$, and each lineage branches with probability $\frac{\alpha}{\alpha i + \theta} = \frac{1}{i + \theta/\alpha}$. Thus, conditional on $A_\tau^\mathcal{K} = m$, the distribution on \mathcal{K}_τ is given by a Chinese

Restaurant Process [2], starting with m tables each with 1 person, and with new tables founded with parameter θ/α .

Let $(x)_{i\uparrow} = x(x+1)\cdots(x+i-1)$ denote the rising factorial. If there is a single mutation with k descendants, then there are $\binom{n-m}{k}$ ways to pick which of the $n-m$ events involve mutant lineages. The probability of a particular such ordering is

$$\frac{\theta (1)_{k\uparrow} (m)_{n-k-m\uparrow}}{\alpha (m + \theta/\alpha)_{n-m\uparrow}} = \frac{\theta (k-1)!(n-k-1)!/m!}{\alpha (n-1)!/m!} + o(\theta).$$

Summing over all $\binom{n-m}{k}$ orderings, and dividing by $\frac{\theta}{2}$, yields

$$f_n^\tau(k \mid A_\tau^\mathcal{K} = m) = \frac{2}{\alpha} \binom{n-m}{k} \frac{(k-1)!(n-k-1)!/m!}{(n-1)!/m!}.$$

References.

- [1] AL-MOHY, A. H. and HIGHAM, N. J. (2011). Computing the Action of the Matrix Exponential, with an Application to Exponential Integrators. *SIAM Journal on Scientific Computing* **33** (2) 488–511.
- [2] ALDOUS, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, (P. L. Hennequin, ed.). *Lecture Notes in Mathematics* **1117** 1–198. Springer Berlin Heidelberg.
- [3] BEAUMONT, M. A. and NICHOLS, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **263** 1619–1626.
- [4] BHASKAR, A., KAMM, J. A. and SONG, Y. S. (2012). Approximate sampling formulae for general finite-alleles models of mutation. *Advances in Applied Probability* **44** 408–428. (PMC3953561).
- [5] BHASKAR, A., WANG, Y. X. R. and SONG, Y. S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research* **25** 268–279.
- [6] BOYKO, A. R., WILLIAMSON, S. H., INDAP, A. R., DEGENHARDT, J. D., HERNANDEZ, R. D., LOHMUELLER, K. E., ADAMS, M. D., SCHMIDT, S., SNINSKY, J. J., SUNYAEV, S. R. et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics* **4** e1000083.
- [7] BRYANT, D., BOUCKAERT, R., FELSENSTEIN, J., ROSENBERG, N. A. and ROY-CHOUDHURY, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* **29** 1917–1932.
- [8] CHEN, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology* **81** 179–195.
- [9] CHEN, H. (2013). Intercoalescence Time Distribution of Incomplete Gene Genealogies in Temporally Varying Populations, and Applications in Population Genetic Inference. *Annals of Human Genetics* **77** 158–173.
- [10] COOLEY, J. W. and TUKEY, J. W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation* **19** 297–301.

- [11] COVENTRY, A., BULL-OTTERSON, L. M., LIU, X., CLARK, A. G., MAXWELL, T. J., CROSBY, J., HIXSON, J. E., REA, T. J., MUZNY, D. M., LEWIS, L. R. et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications* **1** 131.
- [12] DURRETT, R. (2008). *Probability Models for DNA Sequence Evolution*, 2nd ed. Springer, New York.
- [13] EXCOFFIER, L., DUPANLOUP, I., HUERTA-SÁNCHEZ, E., SOUSA, V. C. and FOLL, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* **9** e1003905.
- [14] FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17** 368–376.
- [15] FU, Y. X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology* **48** 172–197.
- [16] GAZAVE, E., MA, L., CHANG, D., COVENTRY, A., GAO, F., MUZNY, D., BOERWINKLE, E., GIBBS, R. A., SING, C. F., CLARK, A. G. et al. (2014). Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences* **111** 757–762.
- [17] GRAVEL, S., HENN, B. M., GUTENKUNST, R. N., INDAP, A. R., MARTH, G. T., CLARK, A. G., YU, F., GIBBS, R. A., BUSTAMANTE, C. D., ALTSHULER, D. L. et al. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108** 11983–11988.
- [18] GRIFFITHS, R. C. and TAVARÉ, S. (1998). The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models* **14** 273–295.
- [19] GUTENKUNST, R. N., HERNANDEZ, R. D., WILLIAMSON, S. H. and BUSTAMANTE, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics* **5** e1000695.
- [20] HIGHAM, N. J. (2002). *Accuracy and Stability of Numerical Algorithms*, 2nd ed. SIAM: Society for Industrial and Applied Mathematics.
- [21] HOPPE, F. (1984). Pólya-like urns and the Ewens’ sampling formula. *J. Math. Biol.* **20** 91–94.
- [22] JENKINS, P. A., MUELLER, J. W. and SONG, Y. S. (2014). General triallelic frequency spectrum under demographic models with variable population size. *Genetics* **196** 295–311. (PMC3872192).
- [23] JENKINS, P. A. and SONG, Y. S. (2011). The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology* **80** 158–173. (PMC3143209).
- [24] JOHNSON, N. L. and KOTZ, S. (1977). *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. Wiley New York.
- [25] KIMURA, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences* **41** 144–150.
- [26] KIMURA, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61** 893.
- [27] KINGMAN, J. F. C. (1982a). The coalescent. *Stoch. Process. Appl.* **13** 235–248.
- [28] KINGMAN, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Prob.* **19A** 27–43.
- [29] KINGMAN, J. F. C. (1982c). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics* (G. Koch and F. Spizzichino, eds.) 97–112. North-Holland Publishing Company.
- [30] LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal*

- of the Royal Statistical Society. Series B (Methodological)* **50** 157–224.
- [31] LUKIĆ, S. and HEY, J. (2012). Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192** 619–639.
 - [32] MORAN, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54** 60–71.
 - [33] NELSON, M. R., WEGMANN, D., EHM, M. G., KESSNER, D., JEAN, P. S., VERZILLI, C., SHEN, J., TANG, Z., BACANU, S.-A., FRASER, D. et al. (2012). An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337** 100–104.
 - [34] NIELSEN, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154** 931–942.
 - [35] PEARL, J. (1982). Reverend Bayes on inference engines: a distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence* 133–136.
 - [36] POLANSKI, A., BOBROWSKI, A. and KIMMEL, M. (2003). A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology* **63** 33–40.
 - [37] POLANSKI, A. and KIMMEL, M. (2003). New Explicit Expressions for Relative Frequencies of Single-Nucleotide Polymorphisms With Application to Statistical Inference on Population Growth. *Genetics* **165** 427–436.
 - [38] SCHAFFNER, S. F., FOO, C., GABRIEL, S., REICH, D., DALY, W. J. and ALTSHULER, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15** 1576–1583.
 - [39] SIDJE, R. B. (1998). Expokit: A Software Package for Computing Matrix Exponentials. *ACM Trans. Math. Softw.* **24** 130–156.
 - [40] STAAB, P. R., ZHU, S., METZLER, D. and LUNTER, G. (2015). scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* btu861.
 - [41] TAVARÉ, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26** 119–164.
 - [42] WAKELEY, J. and HEY, J. (1997). Estimating ancestral population parameters. *Genetics* **145** 847–855.

J. A. KAMM
 DEPARTMENT OF STATISTICS
 UNIVERSITY OF CALIFORNIA, BERKELEY
 BERKELEY, CA 94720
 USA
 E-MAIL: jkamm@stat.berkeley.edu

J. TERHORST
 DEPARTMENT OF STATISTICS
 UNIVERSITY OF CALIFORNIA, BERKELEY
 BERKELEY, CA 94720
 USA
 E-MAIL: terhorst@stat.berkeley.edu

Y. S. SONG
 DEPARTMENT OF STATISTICS AND
 COMPUTER SCIENCE DIVISION
 UNIVERSITY OF CALIFORNIA, BERKELEY
 BERKELEY, CA 94720
 USA
 E-MAIL: yss@stat.berkeley.edu